

MAXIMUM EXPECTED LIKELIHOOD BASED MODEL SELECTION AND ADAPTATION FOR NONNATIVE ENGLISH SPEAKERS

Xiaodong He and Yunxin Zhao

Dept. of Computer Engineering and Computer Science
University of Missouri, Columbia, MO 65211, USA

ABSTRACT

In this paper, the problem of fast model adaptation for nonnative speakers is addressed from a perspective of model complexity selection. The key challenge lies in reliable complexity selection when only a small amount of adaptation data is available. A novel maximum expected likelihood (MEL) based technique is proposed to enable model complexity selection from using as little as one adaptation sentence. In MEL, the expectation of log-likelihood is computed based on the mismatch bias between model and data which is measured by a small amount of adaptation data, and model complexity is selected to maximize EL. Experiments were performed on WSJ data of speakers with a wide range of foreign accents. The proposed method led to consistent and significant improvement on recognition accuracy over MLLR for nonnative speakers, without performance degradation on native speakers. The proposed method was able to dynamically select optimal model complexity as the available adaptation data increased.

1. INTRODUCTION

To improve recognition performance of state-of-art American English speech recognition systems for nonnative speakers remains a challenging task. A straightforward approach is to apply general speaker adaptation techniques such as MLLR and MAP on speaker-independent models to fit the characteristics of a foreign accent. It was shown previously [1] that although speaker adaptation can reduce recognition errors for both native and nonnative speakers, much more adaptation data are needed for the latter than the former to achieve a similar level of performance. A speaker adaptation strategy that focuses on adaptively selecting a proper model complexity for a new speaker has recently been proposed [2,3]. This approach was motivated from the fact that highly detailed English acoustic models with sharp distributions of very narrow allophone classes do not fit well to speech with heavy foreign accents, while a certain level of context-dependent modeling needs to be maintained for discrimination among phones [4]. Therefore, an intermediate level of acoustic model complexity determined from adaptation speech may work best for a foreign accent talker. Among various model complexity selection methods, maximum likelihood (ML) based model selection has been widely used [5]. In the data-rich case, independent “validation data” is employed for model selection. The model that gives maximum likelihood of these data will be selected as the optimal model. However, the require-

ment of large amount of data by ML-based model complexity selection prevents its application in on-line fast speaker adaptation. In [3], an algorithm that combines ML-based state-tying with a pseudo-likelihood (PL) based state-tying was proposed to dynamically determine complexity of acoustic models trained by native English speech for nonnative English speakers by using a small amount of adaptation data. The mismatch between a speaker and an acoustic model was represented by a global bias, which was estimated by using phonetic decision trees of Gaussian Densities (GD). The global bias was then used to compute a PL value for each GD of each tree node, and ML/PL based tree pruning is performed to complete model selection. In [3], although a significant improvement was observed from model selection for recognizing speech of talkers with heavy foreign accent, there were some drawbacks. First, since speech recognition was performed by Gaussian Mixture Density (GMD) based acoustic model, the model selection result obtained through phonetic decision trees of GDs is not precise enough. Second, single global bias is not adequate to comprehensively characterize the detailed mismatch between a speaker’s speech and the phone models.

In the current paper, an expected likelihood (EL) based model selection algorithm is proposed for more accurate model selection, and a comprehensive experimental evaluation is reported on a wider range of accents. The algorithm mainly consists of three steps. First, all allophone states are hierarchically organized through phonetic decision trees and a binary clustering “super” tree, and each node of each phonetic decision trees corresponds to a tied allophone state. For each decision tree node, a GMD is estimated. Second, given a certain amount of adaptation data, the biases between the sample data means and the model means are calculated for each Gaussian component (GC) of each terminal tree node that has adaptation data, and an expected log-likelihood is computed for each node based on the bias distribution. Finally, a bottom-up tree pruning is carried out to select the optimal model complexity that maximizes expected log-likelihood (MEL) over the tree nodes.

This paper is organized as four sections. In section 2 the proposed method is described. In section 3, experimental results are provided. A conclusion is made in section 4.

2. MEL BASED MODEL SELECTION

2.1. Expectation of log-likelihood

In certain scenario of speaker adaptation, only a small amount of adaptation data is available. In such cases, the matching degree

between acoustic model and adaptation data cannot be reliably measured at each tree node by likelihood. However, the matching degree of model and data can be estimated through expected likelihood, which is computed from a small amount of adaptation data based on the GMDs at the tree nodes.

Given a Gaussian mixture density λ , and an arbitrary data set

$\hat{X} = \{x_1, x_2, \dots, x_N\}$, the log-likelihood of \hat{X} is computed as:

$$L(\hat{X} | \lambda) = \sum_{i=1}^N \ln \left[\sum_{k=1}^K w_k \cdot N(x_i; \mu_k, \Sigma_k) \right]$$

where GMD λ has K components, and w_k, μ_k, Σ_k are the weight, mean vector and covariance matrix of the k -th Gaussian component, respectively.

If the likelihood based on the dominant GC is used to approximate the GMD likelihood, the log-likelihood becomes:

$$L(\hat{X} | \lambda) \cong \sum_{k=1}^K \sum_{j \in S_k} \ln[w_k N(x_j; \mu_k, \Sigma_k)]$$

where S_k is the indices set of data that are assigned to the k -th GC with $N_k = |S_k|$.

Assuming $\mu_k = [\mu_{k,1}, \dots, \mu_{k,D}]^T$, $\Sigma_k = \text{diag}[\sigma_{k,1}^2, \dots, \sigma_{k,D}^2]$, then

$$L(\hat{X} | \lambda) \cong -\frac{1}{2} \sum_{k=1}^K \sum_{d=1}^D [N_k \ln(2\pi\sigma_{k,d}^2) + \frac{1}{\sigma_{k,d}^2} \sum_{j \in S_k} (x_{j,d} - \mu_{k,d})^2] + \sum_{k=1}^K N_k \ln(w_k)$$

Denote the mean vector of the data assigned to the k -th GC by $\bar{X}_k = \frac{1}{N_k} \sum_{j \in S_k} x_j$. Let $v_{k,d}^2 = \frac{1}{N_k} \sum_{j \in S_k} (x_{j,d} - \bar{x}_{k,d})^2$ and $b_{k,d} = \bar{X}_{k,d} - \mu_{k,d}$ are the sample variance and the bias between the model mean and the sample mean of $\hat{X}_k = \{x_j, j \in S_k\}$, at the d -th dimension, so

$$L(\hat{X} | \lambda) \cong -\frac{1}{2} \sum_k N_k \sum_d [\ln(2\pi\sigma_{k,d}^2) + \frac{v_{k,d}^2}{\sigma_{k,d}^2} + \frac{b_{k,d}^2}{\sigma_{k,d}^2}] + \sum_k N_k \ln(w_k)$$

The expectation on log-likelihood over sample-data statistics is:

$$\begin{aligned} E[L(\hat{X} | \lambda)] = & -\frac{1}{2} \sum_k N_k \sum_d (\ln(\sigma_{k,d}^2) + \ln(2\pi)) - \frac{1}{2} \sum_k N_k \sum_d \frac{E(v_{k,d}^2)}{\sigma_{k,d}^2} \\ & - \frac{1}{2} \sum_k N_k \sum_d \frac{E(b_{k,d}^2)}{\sigma_{k,d}^2} + \sum_k N_k \ln(w_k) \end{aligned}$$

Assume that the variance of the data is proportional to the variance of the model, i.e., $E(v_{k,d}^2) = \text{const} \cdot \sigma_{k,d}^2$, $\forall k, d$. Also

assume that the number of feature data assigned to the k -th GC is proportional to the weight of that GC, i.e., $N_k = Nw_k$, then we get

$$\begin{aligned} E[L(\hat{X} | \lambda)] = & -\frac{1}{2} N \left[\sum_k w_k \sum_d \ln(2\pi\sigma_{k,d}^2) + \text{const} \cdot D + \sum_k w_k \sum_d \frac{E(b_{k,d}^2)}{\sigma_{k,d}^2} \right] \\ & + N \sum_k w_k \ln(w_k) \end{aligned} \quad (1)$$

2.2. Derivation of bias distribution

To compute the expectation of log-likelihood by Eq. (1), we need to estimate $E(b_{k,d}^2)$. The bias $b_{k,d}$ is viewed as a Gaussian r.v. and the parameter estimation problem is discussed below. For simplicity of notation, the feature dimension index d is omitted in the sequent discussions.

A part of a phonetic state-tying tree is shown in Figure 1. The ‘‘Terminal GC’’ denotes a Gaussian component of a GMD at a terminal node. The ‘‘Full Terminal GC’’ denotes a terminal GC that is assigned adaptation data by Viterbi forced alignment, and

the ‘‘Full internal node’’ denotes a node that covered more than a certain number of ‘‘Full Terminal GC’’ under its sub-tree.

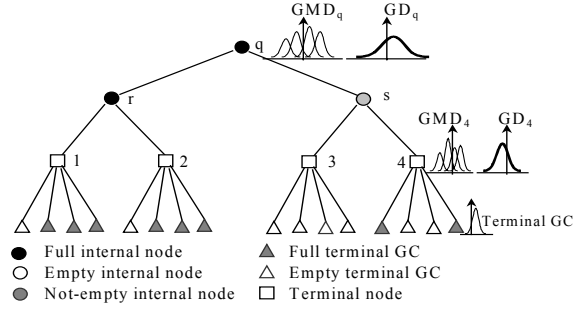


Figure 1: Illustration of the state-tying tree

In the training stage, a GMD and a GD can be generated for each tree node. First, consider GMD and GD of a same node q , where $GMD_q = \sum_{k=1, \dots, K} w_{q,k} N(\mu_{q,k}, \sigma_{q,k}^2)$, $GD_q = N(\mu_q, \sigma_q^2)$. The respective biases are denoted by $b_{q,k}^{(GMD)} = \bar{X}_{q,k} - \mu_{q,k}$ and $b_q^{(GD)} = \bar{X}_q - \mu_q$. As such, $b_q^{(GD)}$ can be approximated as a weighted linear combination of $b_{q,k}^{(GMD)}$, i.e., $b_q^{(GD)} \approx \sum_{k=1, \dots, K} w_{q,k} b_{q,k}^{(GMD)}$. Assume that $b_{q,k}^{(GMD)}, k=1, \dots, K$ are i.i.d. and obey a Gaussian distribution $N(e_q^{(GMD)}, s_q^{(GMD)^2})$. Then $b_q^{(GD)}$ also obeys a Gaussian distribution $N(e_q^{(GD)}, s_q^{(GD)^2})$, with

$$e_q^{(GD)} = e_q^{(GMD)}, \quad s_q^{(GD)^2} = s_q^{(GMD)^2} \sum_{k=1, \dots, K} w_{q,k}^2.$$

On the other hand, refer to Figure 1 and consider the relation between an internal node q and all the terminal nodes $i, i=1, \dots, 4$ below it. We have $b_i^{(GD)} = \bar{X}_i - \mu_i$, and $b_q^{(GD)} = \sum_{i=1-4} \alpha_i b_i^{(GD)}$,

where $\alpha_i = N_i/N_q$ corresponds to the data proportion of node i and node q , $\sum_{i=1-4} \alpha_i = 1$. Consider the relation between the GMD and GD of a same node drawn above, we have

$$b_q^{(GD)} = \sum_{i=1, \dots, 4} \alpha_i b_i^{(GD)} = \sum_{i=1, \dots, 4, k=1, \dots, K} \alpha_i w_{i,k} b_{i,k}^{(GMD)}.$$

For a local sub-tree with root q , we can assume that the bias of the GCs of its terminal nodes, $b_{i,k}, i=1, \dots, 4, k=1, \dots, K$, are i.i.d. and obey a Gaussian distribution $N(e_{termi}^{(GMD)}, s_{termi}^{(GMD)^2})$. Therefore

$$e_q^{(GD)} = e_{termi}^{(GMD)}, \quad s_q^{(GD)^2} = s_{termi}^{(GMD)^2} \sum_{i,k} (\alpha_i w_{i,k})^2.$$

$$\text{Then, } E(b_{q,k}^2) = E(b_{q,k}^{(GMD)^2}) = e_q^{(GMD)^2} + s_q^{(GMD)^2} \quad (2)$$

where

$$e_q^{(GMD)} = e_q^{(GD)} = e_{termi}^{(GMD)}, \quad s_q^{(GMD)^2} = s_{termi}^{(GMD)^2} \cdot \sum_{i,k} (\alpha_i w_{i,k})^2 / \sum_k w_{q,k}^2.$$

The parameters $(e_{termi}^{(GMD)}, s_{termi}^{(GMD)^2})$ can be estimated from the bias samples of GCs of terminal nodes computed from adaptation data. The details are discussed in the next section.

In implementation, we further assume that all terminal nodes have the same number of feature data, which means $N_i = C$. As the result, for a node q , Eqs. (1) and (2) finally become

$$\begin{aligned} E[L(\hat{X} | \lambda_q)] = & -\frac{1}{2} T_q \cdot C \left[\sum_k w_{q,k} \sum_d \frac{E(b_{q,k,d}^2)}{\sigma_{q,k,d}^2} + \sum_k w_{q,k} \sum_d \ln(2\pi\sigma_{q,k,d}^2) \right] \\ & - \frac{1}{2} T_q \cdot C \cdot \text{const} \cdot D + T_q \cdot C \sum_k w_{q,k} \ln(w_{q,k}) \end{aligned} \quad (3)$$

where T_q is the number of terminal nodes under the node q , and

$$E(b_{q,k,d}^2) = e_{q,d}^{(GMD)^2} + s_{q,d}^{(GMD)^2} \quad (4)$$

where

$$e_{q,d}^{(GMD)} = e_{q,d}^{(GD)} = e_{termi,d}^{(GMD)}, \quad s_{q,d}^{(GMD)^2} = s_{termi,d}^{(GMD)^2} \sum_{i,k} (w_{i,k}/T_q)^2 / \sum_k w_{q,k}^2$$

2.3. Estimation of parameters of the bias distribution

For each phone state, a state-tying phonetic-decision tree is built as in [6]. Each node of a decision tree corresponds to a tied state of allophones, and the root node corresponds to a state of a context independent monophone. These monophone states are further organized hierarchically by a binary “super” tree through a clustering procedure. At the beginning, all monophone states are grouped together at the root of the “super” tree. Then binary-split K-means clustering is performed to divide the monophone states into two new children nodes. This procedure continues until each node has only one monophone state. Mahalanobis distance is used in the K-means clustering.

As shown in Figure 1, biases are first computed from the terminal GCs with adaptation data. These biases can be viewed as samples of the bias distribution. As is assumed before, biases of terminal GCs of a local sub-tree can be modeled as i.i.d. Gaussian r.v.’s. Then for a full internal node that has enough samples of bias under it, a distribution $N(e_{termi}^{(GMD)}, s_{termi}^{(GMD)^2})$ of bias can be estimated. In computing of the expected log-likelihood of a node, if it is a full internal node, Eqs. (3) and (4) are applied directly to obtain the EL; otherwise, the distribution of the terminal GC biases under that node is approximated by retrieving the distribution from its nearest ancestor full node.

2.4. MEL approach for model selection

Maximum expected likelihood based model selection attempts to determine an optimal level of model complexity that yields maximum expected log-likelihood. The MEL procedure for model complexity selection is illustrated in Figure 2. Denote expected log-likelihood of a node P by $EL(P)$. The difference between $EL(P)$ and the sum of its two children’s MEL values is:

$$\Delta EL(P, L, R) = [MEL(L) + MEL(R) - EL(P)]/C$$

If $\Delta EL(P) < 0$, then the two children nodes will be pruned. The MEL value of node P is assigned as:

$$\begin{cases} MEL(P) = MEL(L) + MEL(R), & \text{if } \Delta EL(P) > 0 \\ MEL(P) = EL(P), & \text{if P is a terminal node, or } \Delta EL(P) \leq 0 \end{cases}$$

This procedure is carried out bottom-up over all the nodes of a decision tree.

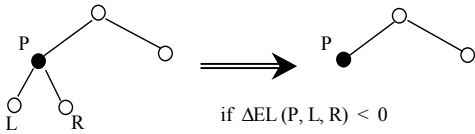


Figure 2: MEL-based tree pruning

With the increasing amount of adaptation data, model parameters can be better adapted and the mismatch bias between a speaker and the model will be reduced. Consequently, the optimal model structure will change with the amount of adaptation data. To dynamically select optimal model, it is desirable to perform model selection after model adaptation. However, the “validation data” used for model selection should be independent with the “training data” used in model adaptation to avoid the over-fitting effect. Taking into consideration of these factors, the complete MEL based model selection/adaptation algorithm is implemented in eight steps as shown in Figure 3.

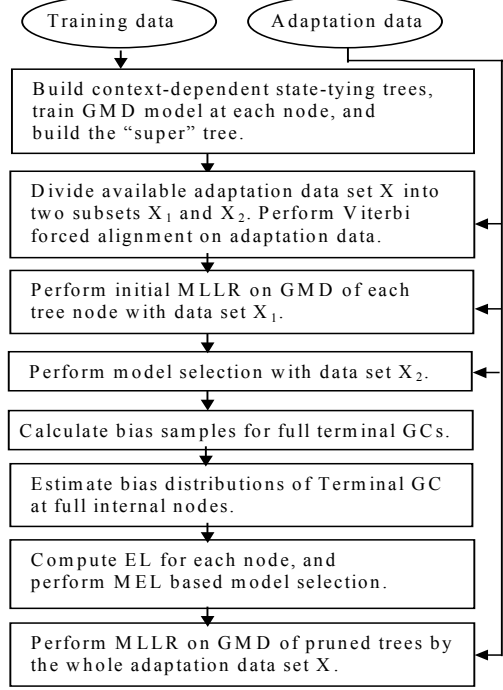


Figure 3: Procedure of model selection/adaptation

3. EXPERIMENTAL RESULTS

3.1. Experimental condition

The proposed method was evaluated on the LDC WSJ1.0 database. The entire set of speaker-independent short-term training data (SI_TR_S, 200 speakers) of WSJ was used for acoustic model training. Each model had three emitting states (except for a “short-pause” model, which had a single state), and each state had a mixture of 16 Gaussian densities. Only internal-word triphones were used. Speech features consisted of 39 components of 12 MFCCs, energy, and their delta and acceleration derivatives. Cepstral Mean Normalization (CMN) as implemented in HTK was applied to both training and test data. Silence model was not adapted.

The state-tying phonetic decision trees were generated by HTK 2.2. For the baseline system, about 103K Gaussian densities after state tying was used. In testing, the standard 5K-vocabulary bigram language model provided with WSJ1.0 was used. The baseline system was tested on WSJ HUB2, where an accuracy of 90.33% was achieved by using the models with 6473 states, which corresponds to phonetic decision trees with totally 6473 terminal states. In the MEL based model selection/adaptation, more detailed decision trees with totally 7137 terminal nodes were generated as the basic trees.

To evaluate the effectiveness of the proposed method, testing was conducted on speakers with different levels of English speaking proficiency. A total of 32 speakers were included in the test set. WSJ database provides two groups of nonnative speakers (DT_S3 and ET_S3) and one group of native speakers (ET_H2), each of which has 10 speakers. In addition, speech data of two speakers with Mandarin Chinese accent (chn1 and chn2) were collected under a similar acoustic condition and with similar prompting texts as WSJ. As show in Table 1, these 32 speakers

were divided into four groups based on their English speaking proficiency.

Group ID	G1	G2	G3	G4
Proficiency	(worst)	(bad)	(good)	(best)
Speakers*	4n0,1,3,4, chn1,chn2	ET_S3 (4nd~4nn)	4n5,8, 9,a,b,c	ET_H2 (4oa~4oj)
Avg. of WER	60.8	26.5	18.1	9.7
Std. of WER	3.8	5.9	7.0	5.9

*4n0,1,3,4,5,8,9,a,b,c belonged to DT_S3

Table 1: Speaker subsets by baseline recognition error rate

In testing, the HTK decoder was used. The decoding parameters, including language model score scale and beam-search pruning thresholds, were optimized for native speaker group ET_H2 and were applied to all the four groups. For each test speaker, the first N adaptation sentences in the adaptation set were used as adaptation data, where N = 1, 5, 10, 20, 40, and the first 20 testing sentences were used in testing (except for the ET_H2 group, where each speaker had only about 20~23 sentences and therefore all the test sentences were used). Recognition results were averaged over each group. In MLLR implementation, Viterbi alignment was used to assign each frame of speech feature to exactly one Gaussian component. The sample size threshold for estimating a MLLR transformation was set to 500, and only mean vectors of Gaussians were adapted. The threshold on number of biases for full node was set to 25.

For different amount of adaptation data, the numbers of utterances for model selection and for initial model adaptation were empirically determined. When the adaptation data were 20 sentences or more, one half of the sentences were used for initial model adaptation and the other half for model selection; otherwise, half of them were used to estimate one single global MLLR transformation and all of the sentences were used for model selection. For the case of one adaptation sentence, only model selection was performed.

3.2. Experimental results

The experimental conditions include MLLR alone (MLLR) and the proposed MEL based method. These results are summarized by word error rate (WER) in Tables 2 through 5.

# adapt. sent.	1	5	10	20	40
MLLR	56.11	41.47	36.14	31.29	27.11
MEL	48.38	32.90	30.30	28.11	25.80
Err. Reduction	13.8	20.7	16.2	10.2	4.8

Table 2: Performance on G1 (WER %)

# adapt. sent.	1	5	10	20	40
MLLR	29.46	20.13	17.54	15.92	16.01
MEL	24.20	18.32	16.99	14.97	13.91
Err. Reduction	17.9	9.0	3.1	6.0	13.1

Table 3: Performance on G2 (WER %)

# adapt. sent.	1	5	10	20	40
MLLR	19.73	13.83	13.53	12.75	11.30
MEL	18.45	12.71	13.29	12.05	11.05
Err. Reduction	6.5	8.1	1.8	5.5	2.2

Table 4: Performance on G3 (WER %)

# adapt. sent.	1	5	10	20	40
MLLR	12.18	9.62	9.24	8.86	8.25
MEL	11.88	9.47	9.84	8.60	8.31
Err. Reduction	2.5	1.6	-6.4	2.9	0.7

Table 5: Performance on G4 (WER %)

The recognition results show that MEL based model selection produced a significant impact on recognition performance for heavy foreign accent speakers. This verified the notion that detail model that is optimal for native speakers is not suitable for heavy accent speakers. Instead, less complex model structures will better tolerate pronunciation deviation of nonnative speech from native English. The error reduction by MEL is observed to reduce with the increase of English speaking proficiency. The MEL based model selection almost did not produce improvement for native speakers. Due to the fact that the model structure of baseline system was optimized for native speakers, and further model selection would not help much.

#Adapt. Sent.	1	5	10	20	40
G1	2124	2352	2726	3187	3221
G2	2614	2546	2831	3384	3397
G3	2665	3001	3194	3879	3569
G4	3425	3264	3902	4464	4410

Table 6: Remaining number of states after model selection

In Table 6, the number of states remained after model selection is shown for each group. We can observe that for speakers with heavy accent, a simpler model structure was selected than that for speakers with slight accent. On the other hand, when more adaptation data became available, more complex models were selected. The proposed MEL algorithm is able to catch this information and dynamically select a more complex model structure. It is also worth noting that, even for native speakers, the MEL selected model is less complex than the baseline system without performance degradation. This is because that, the baseline system is a speaker independent one and it has certain redundancy for individual native speakers.

4. CONCLUSION

Model complexity selection methods as applied to detailed acoustic models in general require large amounts of data in order to reliably compute the fitness of model to data and therefore select a proper level of model complexity. However, the requirement for a large amount of adaptation data is impractical for fast on-line speaker adaptation. In this paper, a novel technique is proposed that perform model selection based on the maximum expected likelihood (MEL) from a small amount of adaptation data. Experimental results indicate that on nonnative English speech, the proposed model complexity selection method led to consistent and significant improvements to MLLR, while the similar recognition performance is maintained for native English speech.

REFERENCES

- [1] Zavalakos, G. Schwartz, R. and Makhoul, J., "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition," *Proc. ICASSP*, pp. 676-679, 1995
- [2] He, X. and Zhao, Y., "Model Complexity Optimization for Nonnative English Speakers," *Proc. EUROSPEECH*, pp.1461-1464, Scandinavia, Denmark, September 2001.
- [3] He, X. and Zhao, Y., "Fast Model Adaptation and Complexity Selection for Nonnative speakers," *Proc. ICASSP*, to appear, Orlando FL, May 2002.
- [4] Compennolle, D., "Recognizing speech of goats, wolves, sheep and ... non-natives," *Speech Communication*, 35 pp.71-79, 2001.
- [5] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.
- [6] Kershaw, D. et al, "The HTK book", <http://htk.eng.cam.ac.uk/docs/docs.shtml>.